Egeninitierad studie

EKONOMISTYRNINGSVERKET

Rapport

Verksamhetsstyrande och andra anslagsvillkor

Automatisk klassificering av innehåll i regleringsbrev 2005–2023

ESV 2023:42

Publikationen kan laddas ner
från ESV:s webbplats esv.se.

Datum: 2023-09-12
Dnr: 2023-07780
ESV-nr: 2023:42

Rapportansvariga: Kenneth Eliasson och Emma Wallerö

Förord

Denna rapport beskriver ett utvecklingsarbete på Ekonomistyrningsverket (ESV) om automatisk klassificering av innehåll i stora textmaterial. En s.k. språkmodell är tränad i uppgiften att skilja på styrande och icke-styrande anslagsvillkor i samtliga myndigheters regleringsbrev för samtliga tillgängliga år. Arbetet är initierat och genomfört av ESV.

Utredarna Annika Alexandersson, Kenneth Eliasson, Hannes Jacobsson, Sven-Olof Junker och Emma Wallerö har deltagit i arbetet. Enhetschefen Martin Sparr har beslutat i ärendet.

Stockholm
2023-09-12

Martin Sparr

Enhetschef

Kenneth Eliasson

Utredare

Innehåll

Sammanfattning

Regeringen kan styra sina myndigheter på flera sätt och ett av dem är att i regleringsbreven till myndigheterna ange villkor för nyttjandet av anslag eller anslagsposter. Det har dock varit oklart i vilken utsträckning som regeringen har styrt myndigheterna med anslagsvillkor. Denna kunskapslucka har funnits trots att anslagsvillkoren har undersökts tidigare av både Ekonomistyrningsverket (ESV) och andra. I tidigare studier har man nämligen bara tagit för givet att anslagsvillkoren är styrande för kärnverksamheternas innehåll eller omfattning. Men det är enkelt att konstatera att vissa anslagsvillkor inte på något sätt är styrande för myndigheternas kärnverksamheter. I denna studie har vi därför undersökt i vilken utsträckning som anslagsvillkoren faktiskt är styrande för myndigheternas kärnverksamheter och om den styrningen har ökat eller minskat över tid.

Eftersom det är ett omfattande arbete att manuellt samla in data för att besvara denna och liknande frågor har ESV bedrivit ett utvecklingsarbete för automatisk klassificering av innehållet i stora textmaterial. Vi har tränat en s.k. språkmodell (KB-BERT) i uppgiften att skilja på styrande och icke-styrande anslagsvillkor i samtliga myndigheters regleringsbrev för samtliga tillgängliga år (2005–2023). Det manuella arbetet har då kunnat begränsas till att ta fram ett klassificeringsschema med kriterier för styrande och icke-styrande anslagsvillkor, samt en mindre referensdatamängd för träningen av språkmodellen.

Resultatet av undersökningen visar att styrande och icke-styrande anslagsvillkor är ungefär lika många fram till och med 2017. Därefter börjar regeringen styra myndigheterna alltmer med anslagsvillkor.

Resultatet motsäger inte de tidigare studierna. Det bekräftar vad vi tidigare visste om styrning med anslagsvillkor, men nu med bättre inblick i vad anslagsvillkoren består av. Att det språkmodellbaserade resultatet motsvarar och är förenligt med tidigare studier visar också på att automatisk klassificering av innehåll i stora textmaterial fungerar.

1 Inledning

Denna rapport har två teman. Det ena temat är metoder för klassificering av innehåll i texter. Sådan datainsamling är ofta resurskrävande, men med stöd av s.k. språkmodeller kan det arbetet effektiviseras. Det andra temat är regeringens styrning av myndigheter. Regeringen kan styra sina myndigheter på flera sätt och ett av dem är att i myndigheternas regleringsbrev ange villkor för nyttjandet av anslag eller anslagsposter.

De frågor vi avser besvara i denna rapport är dels om det går att träna en s.k. språkmodell att skilja på styrande och icke-styrande anslagsvillkor, och hur bra en sådan automatisk klassificering av textinnehåll fungerar. Dels i vilken utsträckning som anslagsvillkoren är styrande för myndigheternas kärnverksamheter, och om den styrningen har ökat eller minskat över tid.

1.1 Bakgrund och problem

Ekonomistyrningsverket (ESV) har till uppgift att utveckla och förvalta den ekonomiska styrningen av den statliga verksamheten. Det innefattar att ha viss kunskap om det aktuella innehållet i budgetprocessens dokument och annat skriftligt material som har betydelse för budgetprocessen. Varje år läser vi därför regleringsbrev, årsredovisningar, budgetunderlag, särskilda myndighetsrapporter, offentliga utredningar, budgetpropositionen och utskottsbetänkanden.

Ibland läser vi dokumenten översiktligt. I så fall händer det att vi kopierar innehåll som i stunden verkar intressant och som skulle kunna användas i något sammanhang, kanske som goda exempel. Ibland läser vi mer systematiskt och riktar in oss på vissa typer av innehåll i något eller några dokument. Det kan handla om hur myndigheterna redovisar resultat,[1] i vilken utsträckning det finns ett jämställdhetsperspektiv i myndigheternas årsredovisningar och budgetunderlag[2] eller hur regeringens styrning via regleringsbreven har förändrats över tid.[3]

Den traditionella metoden för ESV:s mer systematiska läsningar är manuellt utförd kvantitativ innehållsanalys.[4] Förenklat omfattar den metoden följande fem moment:

1. **Utformning av undersökningens ram och fokus**. I detta inledande moment ingår bl.a. att avgränsa och definiera vad det är för innehåll som ska undersökas,

[1] ESV 2002:14, *Effektiv resultatredovisning*; ESV 2012:27, *Prestationer, volymer och kostnader*; ESV 2016:51, *Att synliggöra myndigheters resultat;* ESV 2020:28, *Genomslaget*.
[2] ESV 2023:35, *Jämställdhetsperspektivet i årsredovisningar och budgetunderlag*.
[3] ESV 2021:18, *Regeringens resultatstyrning av myndigheterna*.
[4] Se t.ex. Krippendorff, Klaus (1980, 2019), *Content Analysis. An introduction to its Methodology,* eller Neuendorf, Kimberly A.(2002), *The Content Analysis Guidebook*.

formulera de frågor som ska ställas till materialet samt bestämma eventuella urval och tidsserier.

2. **Utformning av klassificeringsschema**. Det innebär att man formulerar de kriterier som ska vara uppfyllda för att en innehållslig enhet (avsnitt, stycke, mening, bild, etc.) ska kunna klassas som en viss typ av innehåll. Om föremålet för undersökningen är olika typer av mål (t.ex. inriktningsmål, visionära mål och ”smarta” mål) behöver man ta fram kriterier som anger de utmärkande dragen för de måltyper som ska klassificeras.

3. **Träning i att klassificera samma innehåll på samma sätt**. Även om absolut samstämmighet är en omöjlighet behöver arbetsgruppen nå fram till en tillräckligt god samstämmighet. De inblandade behöver läsa och klassificera en liten del av samma material, jämföra sina resultat med varandra och diskutera hur klassificeringsschemat ska tolkas. Samtidigt utvärderar och justerar de klassificeringsschemat. Dels behöver det bli så representativt för materialet som möjligt, dels behöver det bli så lättbegripligt för alla inblandade som möjligt. Vanligtvis innebär detta flera iterativa omgångar med justerade kriterier och nytt material som klassificeras. Så länge arbetsgruppen anser sig sakna tillräckligt god samstämmighet upprepas detta moment.

4. **Insamling av data enligt det slutligt fastställda klassificeringsschemat**. Det är sällan man läser all text i ett dokument. Ofta använder man strategiska sökord i pdf-filens sökmotor för att komma till de ställen i materialet där det sökta innehållet kan finnas. När man upptäcker en innehållslig enhet som motsvarar kriterierna i klassificeringsschemat registrerar man den som en förekomst. Till slut har man sammanställt hur många gånger som de bestämda klasserna av det sökta innehållet förekommer i materialet; man har fått ihop en kvantitativ datamängd.

5. **Analys** av insamlade data och svar på undersökningens frågor.

Metoden är tidskrävande. Framför allt kan momenten 3 och 4 ta mycket tid i anspråk, men hur lång tid kan variera från fall till fall.

Moment 3 kan ta olika lång tid beroende på arbetsgruppens kunskap i undersökningsämnet och dess tidigare erfarenhet av kvantitativ innehållsanalys. För några medarbetare kan moment 3 inte bara bli en fas för kalibrering av samstämmigheten i gruppen utan också en grundläggande inlärningsfas.

Moment 4 kan ta olika lång tid beroende på materialets omfattning. ESV har i flertalet fall samlat in data från ett urval dokument. Med de ca 220 myndigheterna i

redovisningsorganisationen som bas har vi ibland gjort mycket små urval (t.ex. två myndigheter per departement).[5] När populationen är relativt liten, som t.ex. de 53 JiM-myndigheterna, har vi däremot kunnat göra totalundersökningar.[6] Längre tidsserier har vi i regel valt bort på grund av resursbrist. I ett fall skapade vi dock en tidsserie från 1999 till 2020, men med sjuårsintervaller (1999, 2006, 2013 och 2020).[7]

Det är inte helt tillfredsställande att göra små urval, om man vill uttala sig generellt om en population. Med 95-procentig säkerhetsnivå (vanligast i samhällsvetenskap) behöver ett slumpmässigt urval från en population på 220 myndigheter bestå av minst 141 myndigheter. För ett slumpmässigt urval på t.ex. 22 myndigheter behöver man acceptera en 37-procentig säkerhetsnivå.[8]

1.2 Lösning för bättre och snabbare innehållsanalyser

Av de fem beskrivna momenten i den kvantitativa innehållsanalysen är det framför allt datainsamlingsmomentet (moment 4) som helst borde kunna överlåtas till maskinell hantering. Att klassificera innehåll i text är ett regelstyrt och repetitivt arbete. När det görs av människor blir det snart enformigt och tråkigt. Fel kan uppstå på grund av avtagande skärpa hos dem som gör det. I en maskinell datainsamling uppstår inga fel av det slaget och omfattningen på de andra fel som maskinen gör kan uppskattas med hjälp av testdata. Andra fördelar är att totalundersökningar och kompletta tidsserier kan vara standard, eftersom arbetstiden inte är en begränsande faktor i momentet. Det gör det möjligt att borra djupare i materialet och exempelvis jämföra mindre enheter, såsom myndigheter. Därmed kan kvaliteten i slutsatserna förbättras. Säkerhetsnivån kanske inte blir 100-procentig men den blir i varje fall bättre än 37 procent.

1.3 Därför undersöker vi anslagsvillkor i regleringsbreven

Språkmodeller kan effektivisera arbetet med att klassificera innehållet i stora textmängder. Det är tesen i denna rapport. Men för att vi inte bara ska resonera om den tesen teoretiskt prövar vi den även i praktiken. Det förutsätter dels ett maskinläsbart material, dels ett ämne som är undersökt tidigare och som kan tjäna som referens för jämförelser. Materialet regleringsbrev och ämnet anslagsvillkor har dessa respektive egenskaper. Utöver detta erbjuds också en möjlighet till ökad kunskap om anslagsvillkor.

[5] ESV 2020:28
[6] ESV 2023:35
[7] ESV 2021:18
[8] Beräkningarna kommer från en av de ”sample size calulators” som finns på nätet.

1.3.1 Motiv att välja regleringsbrev

Regleringsbreven finns i statens gemensamma informationssystem Hermes sedan 2003. Där finns regleringsbreven i html-format och tillhörande metadata såsom myndighet, departementstillhörighet och årtal. En fördel med html-formatet är att det är maskinläsbart. Det är också enkelt att rensa materialet på sådana element man kan tänkas inte vilja ha, som t.ex. punktlistor eller tabeller, då de omges av särskilda html-taggar. Förutsättningarna att använda hela detta material, från 2003 till i dag, borde därför vara hyggligt goda. Så är inte fallet med t.ex. myndigheternas årsredovisningar och budgetunderlag. Det är först på senare tid, i och med kraven på tillgänglighetsanpassning, som dessa dokument i pdf-format har potential att innehålla så god digital kvalitet att de obehindrat kan läsas maskinellt.

1.3.2 Motiv att välja anslagsvillkor

I detta försök fokuserar vi på anslagsvillkor i regleringsbreven. Det gör vi av två skäl. Det ena är att anslagsvillkorens utveckling över tid har undersökts tidigare. Dessa tidigare studier kan därför tjäna som referenser för jämförelser. Om den språkmodellbaserade studien av anslagsvillkorens utveckling inte avviker från de tidigare manuellt baserade studierna får vi ett kvitto på att den maskinella datainsamlingen fungerar. Enligt tidigare studier ska bl.a. antalet anslagsvillkor öka över tid.[9]

Den andra anledningen är en möjlighet till ny kunskap om anslagsvillkor. Tidigare studier har utgått från att alla anslagsvillkor är styrande för verksamheten. Emellertid behöver man inte läsa många regleringsbrev förrän man upptäcker att villkor till anslag kan avse t.ex. redovisningstekniska frågor, som helt saknar betydelse för hur verksamheten styrs och bedrivs.

Här är tre exempel på faktiska anslagsvillkor:

1. ”Medlen får användas för kärnverksamhetens behov av mätteknisk eller metrologisk FoU samt kunskapsspridning inom området vid SP.”

2. ”Villkoren för stöd till respektive organisation framgår av regeringens beslut den 13 december 2017 (UD2017/20475/IU).”

3. ”Medel som levereras in på grund av ålagd återbetalningsskyldighet för utgift som belastat anslaget ska redovisas på inkomsttiteln 2811 Övriga inkomster av statens verksamhet.”

[9] Ahlbäck Öberg, Shirin och Wockelberg, Helena (2020), Agency control or autonomy? Government steering of Swedish government agencies 2003–2017. I: *International Public Management Journal,* vol. 24, nr 3/2021, s. 330–349. DOI: 10.1080/10967494.2020.1799889 och ESV 2021:18.

(1) kan klassificeras som ett verksamhetsstyrande villkor på egna meriter, (2) är inget självständigt villkor utan en referens till en serie villkor på annan plats och (3) är inget verksamhetsstyrande villkor utan ett redovisningstekniskt sådant.

Frågan är vilken typ av anslagsvillkor som dominerar, den styrande eller den icke-styrande sorten. För att besvara den frågan, och därmed få en bättre bild av regeringens styrning, behöver de styrande anslagsvillkoren klassificeras på ett mer urskiljbart sätt än i de tidigare studierna.

1.4 Metod och material i korthet

För att effektivisera insamlingen av data prövar vi alltså maskinell klassificering av det textinnehåll i regleringsbreven som avser anslagsvillkor. En liten del av materialet har vi klassificerat manuellt för att skapa en referensdatamängd. Större delen av klassificeringen är sedan överlämnad till en språkmodell, som efter ”coachad” träning på referensdatamängden har sorterat anslagsvillkoren som styrande och icke-styrande.

Studien omfattar regleringsbreven till nästan samtliga myndigheter under regeringen och för samtliga år från 2005[10] till 2023. Materialet består av 58 017 meningar som avser potentiella anslagsvillkor av olika sorter.

En förstudie till denna studie gjordes i mastersuppsatsen *Automatic Classification of Conditions for Grants in Appropriation Directions of Government Agencies*[11]. ESV initierade denna förstudie och tog fram materialet till den men det är rapportens författare, Emma Wallerö, som står för innehållet i den då den är ett examensarbete.

1.5 Rapportens disposition

I denna rapport lägger vi stor vikt vid metoden med språkmodellen, eftersom den både är ny i sammanhanget och drastiskt kan förändra förutsättningarna för innehållsanalyser av stora textmängder. Hela kapitel 2 är därför en beskrivning av hur vi har gått till väga i det avseendet.

I kapitel 3 presenterar vi resultatet av studien. För att analysera det lite mer på djupet knyter vi också an till Ahlbäck Öbergs och Wockelbergs studie *Agency control or autonomy? Government steering of Swedish government agencies 2003–2017*. Vi återanvänder deras datamängd och regressionsmodeller, men vi byter ut deras osorterade anslagsvillkor mot våra sorterade.

[10] Av vissa tekniska skäl kan de digitala regleringsbreven för 2003 och 2004 inte ingå denna tidsserie.
[11] Wallerö, Emma (2022), *Automatic Classification of Conditions for Grants in Appropriation Directions of Government Agencies*. Masteruppsats, Uppsala Universitet. DiVa. http://uu.diva-portal.org/smash/record.jsf?pid=diva2:1679811. Notera att i denna förstudie används uttrycket ”annotering” i stället för klassificering av data i syfte att användas som datamängd för träning av en språkmodell.

2 Beskrivning av språkmodellmetoden

I detta kapitel beskriver vi hur en kvantitativ innehållsanalys av text kan göras med stöd av en språkmodell (KB-BERT).[12] Steg för steg går vi igenom hur vi samlade ihop en manuellt klassificerad mindre datamängd, hur vi tränade språkmodellen och beräkningarna av dess träffsäkerhet. Vi diskuterar även språkmodellens karaktär och resultat.

2.1 Översikt av metoden

En kort översikt av vårt tillvägagångssätt:

1. **Förbearbetning av data:** Vi städade textdata på oönskat material (korta meningar, rubriker etc.) samt delade upp data i enheter att klassificera. Vi valde att dela upp anslagsvillkoren i meningar.
2. **Klassificeringsprocess:** Vi, en grupp på fyra utredare, bestämde klassificeringsmetod, alltså vilka klasser som ska användas, diskuterade och nådde en slutsats rörande ifall klasserna bör vara exkluderande eller inte, samt skrev instruktioner för klassificering. Vi testade sedan klassificeringsmetoden genom att (a) var och en testade att klassificera samma villkor och sedan (b) beräknade samstämmigheten. Vi gick sedan igenom de fall där medarbetare klassificerat olika och försökte att komma överens och (c) uppdatera klassificeringsmetoden därefter. Vi upprepade sedan dessa tre steg till dess att samstämmigheten bedömdes som god nog. Vi klassificerade därefter ett något större antal villkor för att kunna använda dem för att träna vår språkmodell.
3. **Dela upp klassificerad data:** För att träna språkmodellen behövde vi en träningsdatamängd, en valideringsdatamängd och en testdatamängd. Träningsdatamängden bör vara den största datamängden och valideringsdatamängden den minsta. Det totala antalet klassificerade meningar var 781 för träningsdatamängden, 87 för valideringsdatamängden, och 386 för testdatamängden.
4. **Utveckling och träning av språkmodell**: Efter viss efterforskning bestämde vi oss för en särskild arkitektur av språkmodell. Vi utgick från den tidigare förstudien (se avsnitt 1.4), och tränade sedan denna språkmodell med hjälp av vår klassificerade data. Vi använde oss av KB-BERT:s Swedish cased BERT-modell som vi finjusterade med hjälp av vår tränings- och valideringsdata.

[12] Bidirectional Encoder Representations from Transformers (BERT) utarbetades av Google och KB-BERT är en svensk BERT-modell som Kungliga biblioteket har tagit fram.

5. **Utvärdering av språkmodell**: Med hjälp av måtten Recall, Precision, f1-score och Accuracy jämförde vi vår språkmodell mot en baslinje med hjälp av vår testdatamängd.[13]

2.2 Förberarbetning av materialet

Det textmaterial som vi utgick ifrån bestod av ett Excel-dokument uppdelat i alla stycken under rubriken ”Villkor till anslag” för varje regleringsbrev 2005–2022. Grunddokumentet var i html och därför fanns mycket html-kod kvar i materialet.[14]

Allt material gick igenom en särskild förbearbetning. Vi valde att dela upp texten i meningar då vi ansåg att större delar, t.ex. stycken, ofta innehöll flera villkor. Att dela upp materialet i meningar bedömdes som det bästa alternativet för klassificeringsuppgiften.

Extraktionen av meningar baserades på att endast välja meningar som kunde vara potentiella villkor. Alla meningar som var kortare än 4 ord eller 16 tecken kasserades. Även rubriker och tabeller och text inom tabeller togs bort. Html-kod togs även bort. Det sista steget i förbearbetningen bestod i att plocka bort meningar som avsåg Regeringskansliet (RK:s disposition).

Förbearbetningen och extraktionen resulterade i 54 555 meningar. Samma typ av bearbetning gjordes för material från 2023 som resulterade i 3 462 meningar. Materialet i denna studie består alltså av sammanlagt 58 017 meningar.

2.3 Klassificeringsprocess

I klassificeringsprojektgruppen ingick fyra personer. Vi började med att formulera ett klassificeringsschema till stora delar baserat på det som användes i förstudien (se avsnitt 1.4). Vi var dock inte helt nöjda med klasserna utan bestämde oss för att omformulera och testa att uppdatera klassificeringsschemat till vi kände oss nöjda.

Vi använde oss alltså av en iterativ klassificeringsprocess där en iteration såg ut på följande vis:

- Testa klassificeringsschemat genom en klassificeringsrunda (där vi klassificerade var och en för sig).
- Utvärdera samstämmigheten.
- Tillsammans gå igenom de meningar vi hade klassificerat olika och diskutera möjliga ändringar i klassificeringsschemat.

[13] Måtten är gängse på området. Se vidare avsnitt 2.5.
[14] Vi har använt de slutliga regleringsbreven, som innehåller alla ändringar som skett under respektive år. För 2023 har dock den första versionen använts, vilket innebär att resultaten för detta år bör tolkas försiktigt.

– Uppdatera klassificeringsschemat.

Efter den fjärde iterationen var vi nöjda med formuleringarna och samstämmigheten, se det slutliga klassificeringsschemat i tabell 1. Eftersom vissa av beskrivningarna är ganska långa strök vi under vissa nyckelord.

Tabell 1: Klasser för meningar från regleringsbrev

Klass	Beskrivning
A Villkor som är styrande för innehållet i verksamheten	Villkor som rör användning av medel och/eller arbetsprocess. Det refererar inte till någon annan styrning. Det är sig självt nog. När villkoret har både innehålls- och omfattningsstyrning kodas det som A (innehållsstyrning)
B Villkor som är styrande för endast omfattningen av medlen	Villkor som rör storlek på medel (i siffror eller ord), men utan att det har egen innehållsstyrning. Även när villkoret har egen omfattningsstyrning och refererar till innehållsstyrning på övergripande nivå eller tidigare beslut kodas det som B (omfattningsstyrning)
C Villkor som inte är styrande för verksamhet eller omfattning	Det kan vara redovisningstekniska villkor, t.ex. att en viss utgift redovisas på en viss inkomsttitel. Icke-styrande villkor kan vidare vara information, t.ex. var anvisningar för hur något ska avräknas finns att ta del av. I de fall storleken på medlen inte anges kan icke-styrande villkor också handla om vem som tillåts disponera medel och vem som tillåts rekvirera medel av vem. Även när villkoret saknar både egen omfattnings-och innehållsstyrning men refererar till styrning på övergripande nivå eller tidigare beslut kodas det som C (icke styrande)
D Inte ett villkor/går inte att avgöra	Det kan exempelvis vara ofullständiga meningar som saknar verb eller kontext

I de två första rundorna klassificerades meningarna i Excel-tabeller, och i rundorna efter detta klassificerades meningarna i ett egenutvecklat klassificeringsverktyg. Ett problem med att klassificera direkt i Excel-filer var att användarna ville ha tillgång till mer metadata och kontext än vad de först fick. När de sedan presenterades med mer metadata och kontext i Excel-filerna bedömdes den vara svåröverskådlig. Därför valde vi att testa att utveckla en egen klassificeringsapplikation som passade för just vår uppgift, och kunde ge en överskådlig visuell representation av departementstillhörighet, tidigare och efterföljande meningar från samma stycke som den klassificerade meningen, anslagspost m.m.

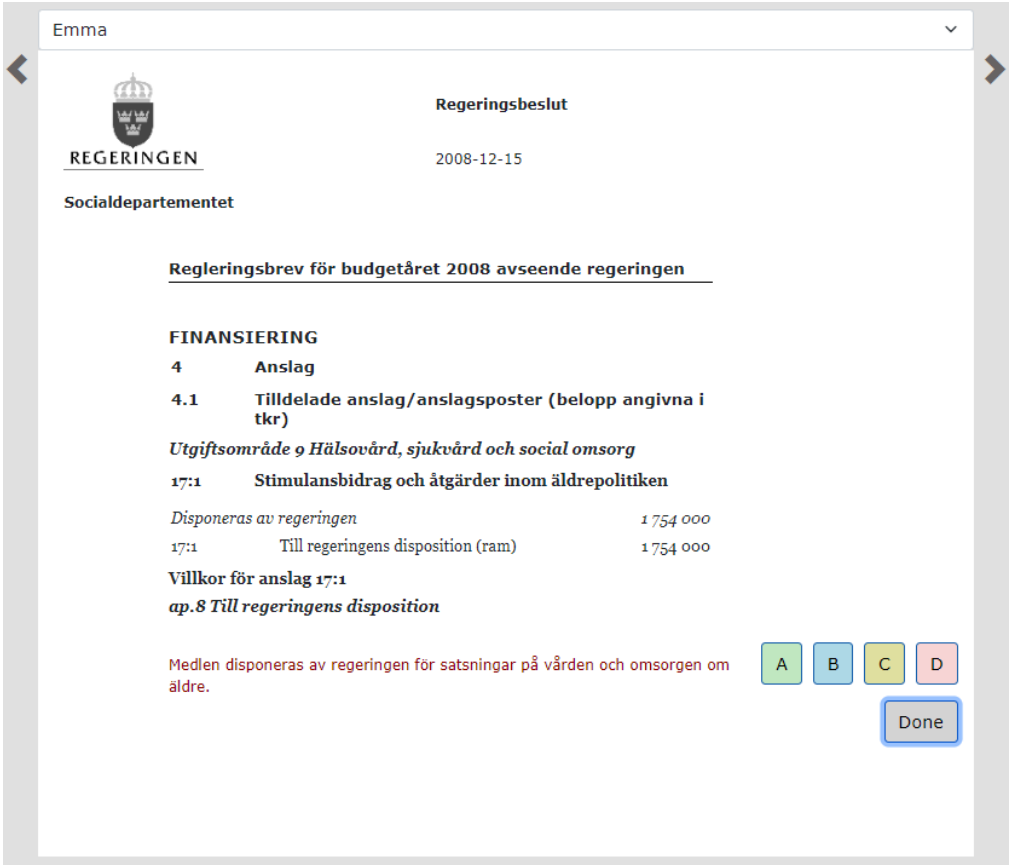
Det finns många verktyg för klassificering men eftersom vi hade ett specifikt behov av presentation och kontext i form av olika sorters metadata (vilket är ovanligt inom klassificering för maskininlärning) passade ingen av dessa för vårt projekt, då de inte

möjliggjorde detta. Vårt egna klassificeringsverktyg tillhandahöll instruktioner, visning av tidigare klassificeringar, samt viss kontext till meningarna (se gränssnitten i figur 1 och figur 2 nedan).

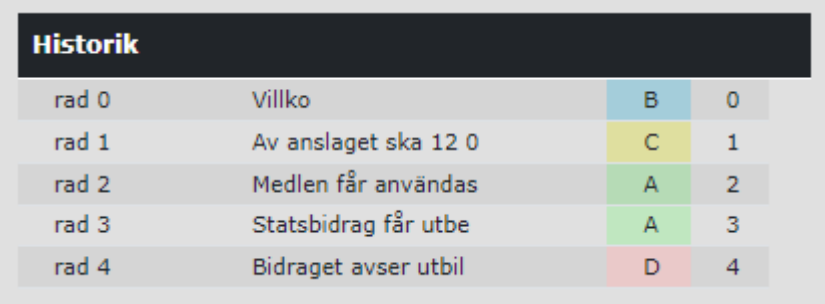
Som figur 1 visar så presenterades anslagsvillkoren som *rödmarkerade* meningar, men i övrigt ungefär som de presenteras i ett regleringsbrev. Medarbetarna skulle sedan välja rätt klass enligt klassificeringsschemat baserat på den rödmarkerade meningen, men även ta viss hänsyn till omgivande aspekter i regleringsbrevet som visades ifall tveksamhet fanns. Klassen angavs genom att välja en av de fyra knapparna A, B, C eller D. Sedan klickade medarbetarna på knappen ”Done”. Då lades den klassificerade meningen till i historiken.

Medarbetarna kunde också alltid gå tillbaka och ändra gamla klassificeringar genom att klicka på de redan klassificerade meningarna som presenterades under ”Historik” (se figur 2). Varje medarbetare kunde dock bara se sin egen historik.

Figur 1: Del av gränssnitt 1



Figur 2: Del av gränssnitt 2



Historik			
rad 0	Villko	B	0
rad 1	Av anslaget ska 12 0	C	1
rad 2	Medlen får användas	A	2
rad 3	Statsbidrag får utbe	A	3
rad 4	Bidraget avser utbil	D	4

Samstämmigheten mättes med hjälp av *Cohen's Kappa koefficient*. Detta mått kan användas för att jämföra klassificeringar av två personer. Måttet bör ses som en indikation snarare än ett exakt mått på överensstämmelse. Kappamått kan befinna sig mellan -1 och 1. Det vanligaste är att få ett mått mellan 0 och 1. Ett kappavärde nära 1 överensstämmer med nästan perfekt samstämmighet.[15]

Efter den 5:e klassificeringsrundan som också var den sista och mest omfattande hade vi nått en samstämmighet på 0,74, 0,70 och 0,48 i kappavärden. Medarbetare 1 och 3 når en måttlig samstämmighet (0,48) medan medarbetare 1 och 2 respektive 2 och 3 når en väsentlig samstämmighet (0,74 och 0,70). Detta var alltså den bästa samstämmigheten vi lyckades nå när vi klassificerade materietalet var och en för sig.

Men för att skapa så bra träningsdata som möjligt ansåg vi att samstämmigheten behövde bli ännu bättre. Därför gick vi tillsammans igenom de meningar där vi hade gjort olika klassificeringar och resonerade oss fram till konsensus. I flera fall var det uppenbart att någon i gruppen hade gjort en miss och att det därför inte krävdes någon ytterligare diskussion för att ändra från en klass till en annan. Det hände också att vi mer ingående behövde diskutera våra skilda klassificeringar. Vi lyckades ändå alltid avsluta varje sådan diskussion i enighet. Det var heller ingen i gruppen som alltid vann dessa diskussioner. Var och en i gruppen gav med sig ungefär lika många gånger.

2.4 Data

I detta avsnitt beskrivs de klassificerade meningar som användes i processen.

2.4.1 Datatyper

Tre typer av data ingår i våra datamängder:

- Typ 1: Manuellt klassificerad data
- Typ 2: Augmenterad data

[15] Sim, Julius och Wright, Chris C. (2005) The kappa statistic in reliability studies: use, interpretation, and sample size requirements. I: *Physical therapy*, vol. 85, nr 3/2005, s. 257–268,

– Typ 3: Automatiskt klassificerad data som är kontrollerad och rättad.

Typ 1 Manuellt klassificerad data

Denna datatyp var manuellt klassificerad data baserad på det slutliga klassificeringsschemat (se tabell 1). För denna datamängd var varje mening klassificerad av två medarbetare. Detta för att kunna kontrollera samstämmigheten och detektera svårtolkade meningar. Tre personer genomförde denna klassificering. Vi gick sedan tillsammans igenom de meningar där medarbetarna hade angett olika klasser och kom överens om vilka klasser dessa meningar bör ha.

Datatyp 1 var skevt fördelad mellan klasserna. Vanligast var A med 42 procent, C med 36 procent, B med 12 procent och D med 9 procent. Det fanns 331 meningar av denna data typ.

Typ 2 Augmenterad data

Då den manuellt klassificerade datatypen var skevt fördelad mellan klasserna fanns det anledning att se över möjligheterna att använda sig av syntetisk data baserad på den lilla manuellt klassificerade datamängd vi redan hade. Den metod som användes kallas "Easy Data Augmentation" (EDA) och består av enklare sätt att augmentera data, exempelvis att byta ut två ord i en mening mot varandra, ta bort ett ord ur en mening, eller ta bort eller flytta på vissa tecken i ord. Att byta ut slumpmässiga ord mot synonymer är ytterligare en metod som ingår i EDA. Detta gjordes dock inte då den kod och korpus som användes var baserad på engelska.[16] Det skapades 280 meningar av typ 2.

Typ 3 Automatiskt klassificerad och rättad data

Denna datatyp bestod av 324 automatiskt klassificerade meningar av en finjusterad (fine-tuned på engelska) språkmodell (KB-BERT) som tränats på datatyperna 1 och 2. Dessa rättades manuellt och användes i tränings- och valideringsdatamängd för senare språkmodell. För ungefär 25 procent av meningarna korrigerades tillhörande klass. Modellen tenderade att ofta klassificera villkor som hör till klass A som C, villkor som hör till klass D som A, och villkor som hör till klass C som A (se tabell 2). Vidare genererades en liknande datamängd på 319 meningar av en senare KB-BERT-modell som finjusterats på datatyp 1, 2 och 3. Även denna datamängd rättades, och lades sedan in i testdatamängden för att komplettera de 67 meningarna av typ 1.

[16] Wei, Jason och Zou, Kai (2019), *Eda: Easy data augmentation techniques for boosting performance on text classification tasks*. arXiv preprint arXiv:1901.11196.

2.4.2 Slutliga datamängder

Tränings och valideringsdatamängd

Träningsdatamängden bestod av 781 klassificerade meningar och valideringsdatamängden bestod av 87 klassificerade meningar.

I tränings- och valideringsdatamängden ingick 264 klassificerade meningar av typ 1, 280 klassificerade meningar av typ 2, samt 324 klassificerade meningar av typ 3 (se tabell 3). Träningsdatamängden bestod till 90 procent av dessa datatyper och valideringsdatamängden utgjorde resterande 10 procent.

Testdatamängd

Testdatamängden bestod av totalt 386 klassificerade meningar.

Testdatamängden består dels av 67 klassificerade meningar av typ 1, dels av 319 klassificerade meningar av typ 3. Dessa data var slumpmässigt valda med en klassdistribution som speglar den totala datamängden av typ 1. Anledningen till att vi inte använde augmenterad data till testdatamängden var för att vi ville använda oss av verkliga anslagsvillkor och inte syntetiska[17] sådana när vi utvärderade modellen.

Tabell 2: Fördelning av antal datapar per datatyp i tränings, validering och testdatamängd

	Typ 1	Typ 2	Typ 3	Totalt
Träningsdatamängd och Valideringsdatamängd	264	280	324	868
Testdatamängd	67		319	386

2.5 Träning av språkmodell

Formen på data-input till språkmodellen består av en mening. Språkmodellen som användes i studien var KB-BERT:s ”BertForSequenceClassification", specifikt ”Cased Swedish BERT". Denna finjusterades med hjälp av vår egen tränings- och valideringsdata.

För att modellen ska kunna bearbeta vårt textmaterial måste man dela upp texten i s.k. tokens. Korta ord blir oftast ett token medan längre ord delas upp i flera tokens. Det kan skilja mellan olika tokeniserare. Tokeniseraren som användes var ”BERT tokenizer” som delar upp text i mindre fragment.[18]

[17] Syntetisk data är alltså data som skapas istället för att plockas från verkliga exempel. Eftersom vår augmenterade data inte kan återfinnas ordagrant i regleringsbrev (eftersom vi modifierat meningarna något för denna datamängd) kan man säga att denna datatyp var syntetisk.

[18] Devlin, Jacob m.fl. (2018), *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint. arXiv:1810.04805.

Varje språkmodell utvärderades med hjälp av relevansmåtten Accuracy, Precision, Recall och f1-score. Dessa begrepp används oftast på engelska i svensk text.[19] Men de kan översättas enligt följande:

- **Accuracy:** Andelen korrekta klassificeringar.
- **Precision:** Andelen korrekta klassificeringar bland alla positiva klassificeringar.
- **Recall:** Andelen korrekta klassificeringar bland alla sanna positiva.
- **F1:** En sammanvägning (det geometriska medelvärdet) av precision och recall.

Det finns flera orsaker till varför olika mått används för att utvärdera en modell. Ibland kan det viktigaste vara att inte missa några positiva exempel, då är recall ett viktigt värde. Ibland är det viktigaste att andelen korrekt klassificerade positiva är hög, då är precision ett viktigare mått. Accuracy är det mest lättolkade, men kan överskatta modellens förmåga när klasserna som ska prediceras är olika stora (att alltid predicera majoritetsklassen kan då ge ett högt värde på accuracy).

2.5.1 Testkörningar

Vi testade att finjustera språkmodellen med olika datamängder under projektets gång.

2.5.1.1 Test 1

Till att börja med tränades en modell endast på den manuellt klassificerade datamängden. Denna typ av modell var instabil och genererade ganska olika resultat jämfört med andra modeller som tränats på samma data och parametrar. Generellt låg resultaten runt en Accuracy på 0,75. För att potentiellt förbättra modellen bestod nästa steg av att göra något åt den ojämna klassdistributionen.

2.5.1.2 Test 2

Sedan tränades en modell utöver manuellt klassificerade datamängden även på den augmenterade datamängden. Denna modell presterade bättre med en Accuracy på närmare 0,88. Modelltypen var dock fortfarande instabil. Resultatet förändrades ganska mycket från en körning till en annan. På dessa data gjordes en s.k. grid search[20] för att optimera parametrarna epok, learning rate samt batch-storlek. Samtliga möjliga kombinationer testades. De bästa parametrarna var 3 epoker, en learning rate på 0,00005 samt en batch-storlek på 16.

[19] Enligt DIGG kan dessa mått tappa delar av sin betydelse om man översätter dem från engelska till svenska. Se DIGG, *Testa ny teknik för automatisering inom offentlig förvaltning* [pdf], https://www.digg.se/download/18.79c61f7c17db5871992f0c0/1647952780135/testa-ny-teknik-for-automatisering-inom-offentlig-forvaltning.pdf. Hämtad 29 augusti 2023.

[20] Grid search innebär att man testar slumpmässiga värden på utvalda parametrar mot varandra. De värden som kan slumpas mellan kan förbestämmas.

Detta var dessa parametrar som användes för denna modell:

- Learning rate: 0,00005, 0,00003, 0,00001
- Batch-storlek: 16, 32
- Epoker: Använde ”Early stopper”.[21]

2.5.2 Slutgiltig språkmodell

Sist tränades en modell på utöver ovan nämnda datamängder också på en bootstrappad datamängd, dvs. data som klassificerats av en språkmodell och sedan rättats.[22] Denna typ av modell är mer stabil och ändrar sig mycket mindre från träning till träning. De parametrar som användes här var 7 epoker, en learning rate på 0,00005 och en batch-storlek på 32. Det är denna modell som vi presenterar resultat för i denna studie, och den som vi använt för att automatiskt klassificera de anslagsvillkor vi presenterar i kapitel 3.

2.5.3 Baslinje

Den finjusterade språkmodellen jämfördes sedan mot en baslinje som förutser alla meningar som den vanligast förekommande klassen i träningsdatamängden. I vår studie var A den vanligaste klassen, och därför klassade baslinjen alla meningar i testdatamängden som A. Denna baslinje kallas i denna studie för majoritetsklass-baslinje.

2.6 Resultat för metod

Den slutliga modellen där träningsdatadatamängden bestod av alla tre datatyper genererade en Accuracy på 0,85, Precision på 0,80, Recall på 0,76 och f1-score på 0,78 (se tabell 4). I tabell 4 finns resultaten från baslinjen där alla meningar förutses som A. Dessa resultat är avsevärt sämre med en Accuracy på 0,46 och en f1-score på 0,16.

Tabell 3: Resultat för språkmodell och baslinje

	Precision	**Recall**	**f1-score**	**Accuracy**
Majoritetsklass-baslinje	0,12	0,25	0,16	0,46
Finjusterad KB-BERT	0,80	0,76	0,78	0,85

Prestandan för varje klass redovisas i tabell 5. Klass A, C och D har generellt goda resultat, medan klass B har ett Recall på 0,61 som sticker ut något. Detta kan bero på

[21] Att använda en ”Early stopper” inom maskininlärning handlar om att inte träna vidare en modell som nått sin max-potential givet de parametrar som anges. Efter varje tränad epok stämmer vi av ifall modellen blivit bättre än den senast tränade epoken. Om inte så stannar ”Early stopper” på den tidigare epok som gav bättre resultat.
[22] Stollenwerk, Felix m.fl. (2022), *Annotated Job Ads with Named Entity Recognition* [pdf], https://2022.sltc.se/papers/SLTC22_paper_3062.pdf. Hämtad 30 augusti 2023.

att testdatamängden är så pass liten och dessutom stratifierad, vilket resulterar i att det finns väldigt få punkter i testdatamängden att utvärdera meningar av klass B mot. Även klass D har få förekomster i testdatamängden.

Tabell 4: Resultat för varje klass för finjusterad KB-BERT

	Precision	Recall	f1-score	Antal observationer
A: innehåll	0,90	0,86	0,88	179
B: omfattning	0,67	0,61	0,64	23
C: ej styrande	0,84	0,91	0,87	152
D: ej villkor	0,78	0,66	0,71	32

Av klass B:s låga Recall att döma kan man misstänka att det finns en del oupptäckta meningar av klass B i förutsägelser av språkmodellen. Detta är bra att tänka på när man i kapitel 3 tolkar resultaten av automatisk klassificering av anslagsvillkor från 2005–2023.

3 Anslagsvillkorens utveckling enligt KB-BERT

I det här kapitlet redogör vi för resultatet av undersökningen. Inledningsvis rekapitulerar vi undersökningens klasser av anslagsvillkor och presenterar ett par grafer över hur dessa klasser utvecklas från 2005 till 2023. Det vi fokuserar på här är om de automatiskt genererade kurvorna över antalet anslagsvillkor ökar, vilket de borde göra enligt de tidigare manuella studierna.

Därefter tar vi hjälp av tidigare forskning i bemärkelsen att vi återanvänder den datamängd och de regressionsmodeller som Öberg Ahlbäck och Wockelberg använde till sin studie.[23] Här tittar vi på om några av regressionskoefficienterna får andra värden när vi lägger in våra maskinellt framtagna data om anslagsvillkor i modellerna. Det är ett ytterligare test på hur väl en automatisk klassificering av innehållet står sig mot en manuell. För att den språkmodellbaserade studien ska klara testet ska klasserna av anslagsvillkor (styrande och icke-styrande samt hopslagna) få ungefär samma regressionskoefficienter som i den tidigare manuella studien med bara osorterade anslagsvillkor. Dessutom kan vi se om det finns skäl att dra nya slutsatser om regeringens samlade styrning.

Till sist återknyter vi till en tidigare ESV-studie om förhållandet mellan anslagsposter och anslagsvillkor. Frågan vi ställer med referens till den studien är om språkmodellens utfall visar på ett liknande resultat.

3.1 Hur verksamhetsstyrande och andra anslagsvillkor har utvecklats

Alla anslagsvillkor är inte styrande för myndigheters verksamhet. Det går att konstatera med bara ett fåtal regleringsbrev som källor. I denna studie används följande indelning av anslagsvillkor samt en övrigkategori för sådant som inte kan tolkas som anslagsvillkor:

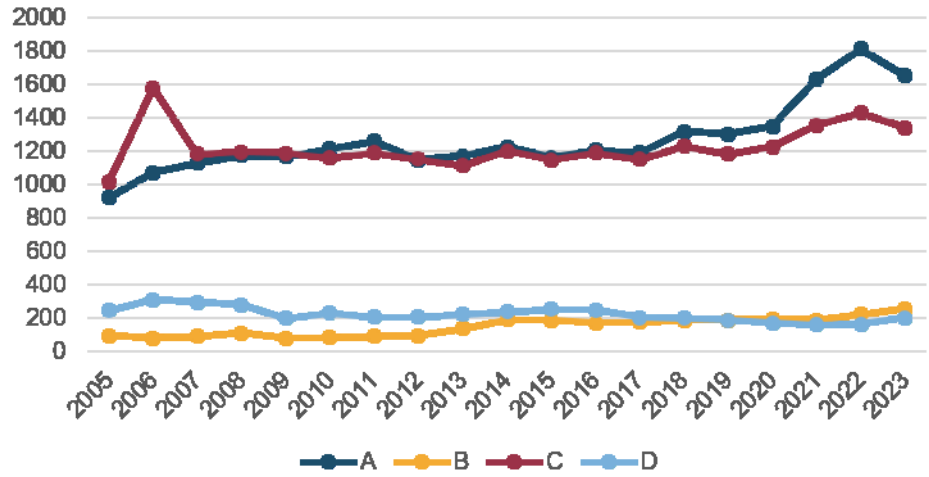
- A: Villkor som är styrande för innehållet i verksamheten.
- B: Villkor som är styrande för endast omfattningen av medlen.
- C: Villkor som inte är styrande för verksamhet eller omfattning.
- D: Inte ett villkor/går inte att avgöra.

Klasserna A–D avser vara heltäckande. Åtminstone kunde anslagsvillkoren fördelas enligt detta schema när vi manuellt tog fram den mindre träningsdatamängen. Figur 3

[23] Materialet finns tillgängligt på Svensk nationell datatjänst, SND-ID: 2022-39-1: Svenska regeringars krav på återrapportering från statliga myndigheter 1993–2017.

visar utfallet av den tränade och finjusterade KB-BERT-modellens klassificering av A–D i samtliga myndigheters regleringsbrev från 2005 till 2023.

Figur 3: Antal anslagsvillkor av varje klass

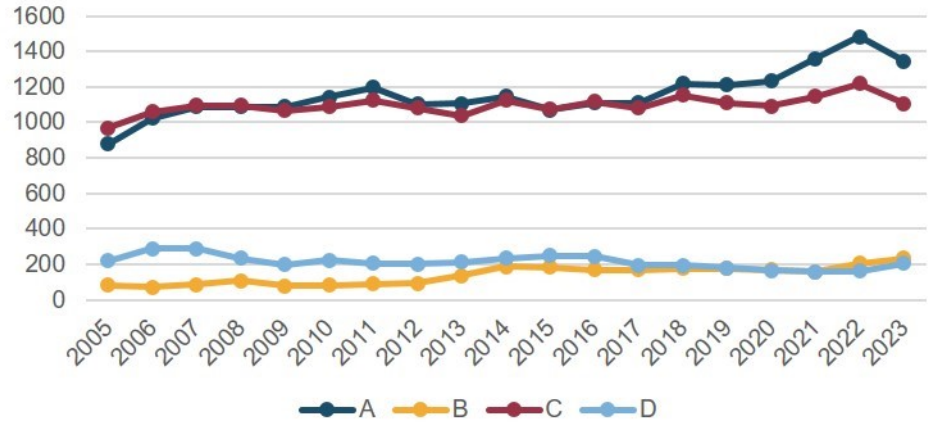


Både styrande (A och B) och icke-styrande villkor (C) visar på en ökning över tid. Trots nedåtgången det sista året är trenden ändå förväntad, enligt hur den borde vara mot bakgrund av tidigare studier. Övrigkategorin (D) samlar relativt få träffar. Den automatiska klassificeringen av anslagsvillkor verkar ha fungerat i den meningen att den har gett ett väntat resultat.

Om vi fokuserar endast på de verksamhetsstyrande och icke-styrande anslagsvillkoren (A respektive C) ser vi ett intressant mönster i materialet. Dessa klasser av anslagsvillkor följs åt och är lika omfattande fram till 2018, då de glider isär. De allra senaste åren blir de verksamhetsstyrande anslagsvillkoren betydligt fler än de icke-styrande. Kurvorna tyder alltså på att regeringen i ökande omfattning styr myndigheternas kärnverksamheter med anslagsvillkor.

Piken 2006 är besynnerlig. Vi kan inte förklara den men vi kan lokalisera den. Den består av icke-styrande villkor i regleringsbreven till universitet och högskolor det året. Vi borrar inte djupare i detta här, men det kan vara ett uppslag för en eventuell kommande fallstudie. Figur 4 visar utvecklingen när universitet och högskolor är borttagna från datamängden.

Figur 4: Antal anslagsvillkor av varje klass utan universitet och högskolor



Den ökande trenden kvarstår även efter det att universitet och högskolor är borttagna. Ett skäl att provisoriskt ta bort dessa myndigheter från materialet är att piken 2006 kan ge missvisande analyser av materialet. Universitet och högskolor utgör en så stor grupp att konstigheter i materialet för den gruppen slår igenom och påverkar slutsatserna för alla andra myndigheter. Variablerna för styrande och icke-styrande anslagsvillkor innehåller därför inte data från universitet och högskolor när vi lägger in dem i regressionsmodellerna i avsnitt 3.2.2 nedan.

3.2 Anslagsvillkor i regeringens samlade styrning av myndigheter

Innan vi ger oss i kast med regressionsmodellerna behöver bakgrunden till dessa beskrivas. Modellerna ingår i Ahlbäck Öbergs och Wockelbergs studie om regeringens myndighetsstyrning och myndigheternas autonomi.[24] Frågan de diskuterar med stöd av dessa modeller är om summan av regeringens samlade styrning är konstant, dvs. om minskad styrning med ett styrmedel kompenseras av ökad styrning med ett annat styrmedel.

Som Ahlbäck Öberg och Wockelberg noterar minskade regeringen sin styrning med återrapporteringskrav strax innan 2010-talets början. Om allt annat var lika skulle utrymmet för myndigheternas autonomi därmed ha blivit större.

3.2.1 Analysen i refererad version

För att analysera om det blev mer autonomi eller inte identifierar Ahlbäck Öberg och Wockelberg en serie andra styrmedel än återrapporteringskrav, som beroende på hur

[24] Ahlbäck Öberg, Shirin och Wockelberg, Helena (2020), ”Agency control or autonomy? Government steering of Swedish government agencies 2003–2017”, i: *International Public Management Journal*, DOI: 10.1080/10967494.2020.1799889.

de används kan vidga eller krympa ramarna för myndigheterna självständiga beslutsfattande.

De andra styrmedlen är:

- graden av anslagssparande (mer eller mindre än 3 procent)
- tilldelad föreskriftsrätt (förekomst)
- organisationsform (styrelsemyndigheter/enrådighetsmyndigheter)
- anslagsvillkor (antal).

När variablerna körs i en regressionsmodell som ska förklara variationen av antalet återrapporteringskrav, blir det ett signifikant samband mellan antalet anslagsvillkor och antalet återrapporteringskrav. Utan någon ytterligare kontrollvariabel i modellen står ett minskat antal återrapporteringskrav i relation till ett ökat antal anslagsvillkor. Men sambandet försvinner vid kontroll för anslagens storlek. Myndigheter med stora anslag har också många anslagsvillkor. Ahlbäck Öbergs och Wockelbergs studie ger därför inget stöd för hypotesen att anslagsvillkor har betydelse för hur regeringen balanserar olika sätt att styra myndigheterna. Ett minskat antal återrapporteringskrav har inte bytts ut mot ett ökat antal anslagsvillkor. Även sambandet mellan återrapporteringskrav och föreskriftsrätt försvinner vid kontroll för anslagens storlek.

Däremot kvarstår sambandet för anslagssparande som är mindre än 3 procent. Detsamma gäller organisationsform. Myndigheter som omvandlas från styrelsemyndigheter till enrådighetsmyndigheter får ett mindre antal återrapporteringskrav. Analysen visar alltså att det minskade antalet återrapporteringskrav har kompenserats av ett mindre utrymme för anslagssparande och en mer direkt kanal till myndighetschefen.

3.2.2 Analysen i återskapad och utvecklad version

Vi gör nu om analysen. Samma data används som i Ahlbäck Öbergs och Wockelbergs regressionsmodeller, men vi byter ut deras osorterade data om anslagsvillkor mot våra sorterade.

Regressionsmodellerna i tabell 6 skiljer sig åt i de bemärkelserna att Modell 1 har hopslagna (styrande och icke-styrande) anslagsvillkor, Modell 2 endast styrande anslagsvillkor, Modell 3 endast icke-styrande anslagsvillkor och Modell 4 styrande anslagsvillkor med kontroll för anslagsstorlek.

Tabell 6: Regressionsmodeller med befintligt och nytt material

	Modell 1	Modell 2	Modell 3	Modell 4
(Intercept)	84,994***	85,354***	85,575***	-216,055***
	(11,823)	(11,824)	(11,831)	(36,805)
>3 % anslagssparande	4,535	3,076	6,590	5,078
	(15,139)	(15,175)	(15,139)	(14,794)
<3 % anslagssparande	-57,613***	-51,565***	-59,523***	-89,111***
	(15,973)	(15,628)	(16,285)	(15,699)
Föreskriftsrätt	39,132***	39,176***	41,586***	-1,024
	(11,661)	(11,694)	(11,572)	(12,191)
Organisationsform	-56,242***	-57,067***	-55,583***	-52,005***
	(11,299)	(11,304)	(11,317)	(10,951)
Anslagsstorlek (log)				26,596***
				(3,075)
Hopslagna villkor (ESV)	0,601***			
	(0,158)			
Styrande villkor (ESV)		1,087***		-0,095
		(0,297)		(0,317)
Icke-styrande villkor (ESV)			1,041***	
			(0,298)	
Antal obs.	933	933	933	927
R^2	0,065	0,064	0,063	0,135

*Kommentar: Beroende variabel i modellerna är antalet återrapporteringskrav. Markeringen "***" visar att sambandet är signifikant på 0,01-nivån. Styrande anslagsvillkor är A och B tillsammans. Icke-styrande anslagsvillkor är C. Övrigkategorin D ingår inte. För anslagssparande är referenskategorin de myndigheter som har precis 3 %.*

I denna analys är vi intresserade av hur koefficienterna uppträder med de nya data om anslagsvillkor som den finjusterade språkmodellen KB-BERT har levererat. I den ursprungliga analysen har anslagsvillkoren innan kontrollen för anslagsstorlek en koefficient på värdet 1,216.[25] I denna utvecklade analys har de styrande anslagsvillkoren en koefficient på 1,087 och de icke-styrande 1,041. Det är snarlika värden som i den ursprungliga analysen. Värdet för de hopslagna villkorstyperna avviker något mer (0,601), men det är ändå i samma härad. Vi konstaterar därför att våra maskinellt framtagna data om sorterade anslagsvillkor i stort sett motsvarar de tidigare manuellt framtagna data om osorterade anslagsvillkor.

Vi kan också konstatera att regressionsmodellerna i vår tappning ger samma övergripande resultat som de ursprungliga. Även här faller anslagsvillkor bort som en signifikant förklaringsfaktor vid kontroll för anslagsstorlek. Samtidigt ger våra nya

[25] Ahlbäck Öberg och Wockelberg (2020), s. 13.

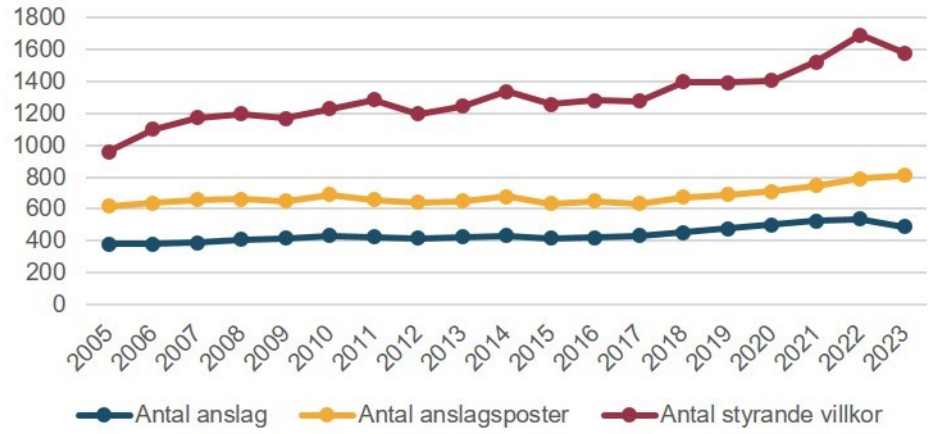
data om anslagsvillkor inget stöd för andra slutsatser om regeringens styrningsbalans än de slutsatser som redan är dragna.

3.3 Anslag, anslagsposter och anslagsvillkor

I ESV-rapporten *Regeringens resultatstyrning av myndigheterna. En kartläggning av instruktioner och regleringsbrev under två decennier* (ESV 2021:18) redovisas bl.a. hur anslagsposter och osorterade anslagsvillkor har utvecklats över tid. Tidsserierna löper från 1999 till 2020 i sjuårsintervaller (1999, 2006, 2013 och 2020). Urvalet myndigheter är extremt litet. Det är en myndighet per departement och de är valda med hänsyn till att de är olika stora. Det är förstås inte möjligt att med någon högre säkerhet hävda att resultatet från den undersökningen är representativt för alla myndigheter. Undersökningen kan egentligen bara avse utvecklingen för de 11 myndigheterna som ingår i urvalet. För dessa 11 myndigheter har både anslagsposter och osorterade anslagsvillkor ökat i antal.

En språkmodellbaserad studie kanske kan fungera som en mer trovärdig grund för mer generella påståenden. Figur 5 visar utvecklingen av anslagsvillkor enligt den finjusterade KB-BERT-modellen och utvecklingen av anslag och anslagsposter enligt ESV:s registerdata.

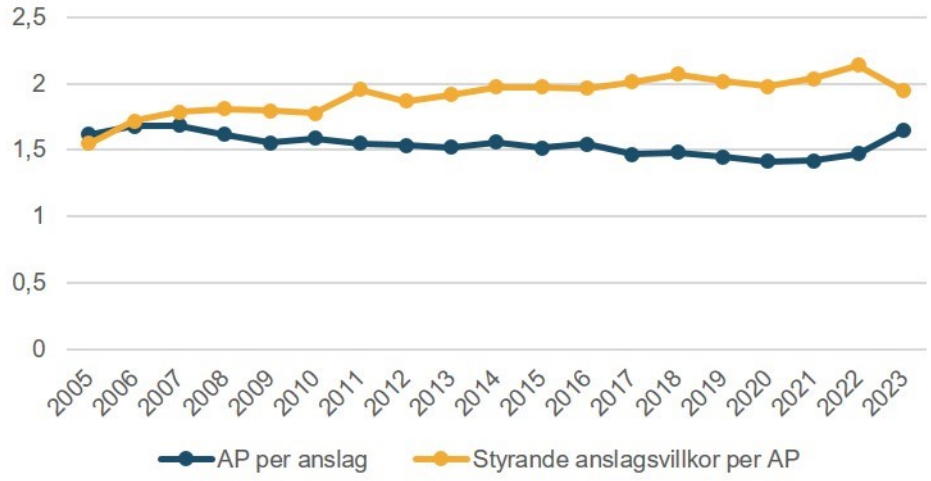
Figur 5: Antal anslag, anslagsposter och styrande anslagsvillkor från 2005 till 2023 (inga universitet eller högskolor)



Kommentar: Kurvan för styrande anslagsvillkor är sammansatt av både klassen för innehållsstyrning (A) och klassen för omfattningsstyrning (B).

Vi kan då se att det blir samma mönster utifrån både den manuella fåtalsstudien och den maskinella totalstudien. Det blir mer av allting: anslag, anslagsposter och styrande anslagsvillkor. Även förhållandet mellan anslagsposter och styrande anslagsvillkor visar på en ökande trend (se figur 6).

Figur 6: Antal anslagsposter per anslag och antal styrande anslagsvillkor per anslagspost (inga universitet eller högskolor)



Det är ungefär 1,5 styrande anslagsvillkor per anslagspost 2005 och ungefär 2 styrande anslagsvillkor per anslagspost 2023. Enligt studien om de osorterade anslagsvillkoren för de 11 myndigheterna är dessa förhållanden 2,5 år 2006 och 2,9 år 2020. Trenden är alltså densamma. De osorterade anslagsvillkoren borde dessutom vara fler per anslagspost än de urskilda styrande anslagsvillkoren, vilket ju resultatet också visar.

ESV konstaterar därmed att utfallet enligt vår finjusterade KB-BERT-modell motsvarar och är förenligt med det tidigare kända resultatet. Regleringsbrevets finansiella del är en alltmer använd kanal för detaljstyrning.

4 Diskussion och reflektion

Det har visat sig vara en svår uppgift att manuellt skilja på styrande och icke-styrande anslagsvillkor. Villkoren är alltför mångskiftat skrivna för att vart och ett av dem ska kunna bli korrekt sorterade utifrån enkla klassificeringsregler som alla inblandade förstår på samma sätt. I studien har vi ändå nått en nivå av samstämmighet som vi finner god nog, och kan därför säga att de klassificeringsregler vi utformat för uppgiften är relativt välfungerande.

Om vi hade haft ännu mer välfungerande klassificeringsregler som resulterat i ännu bättre samstämmighet, så hade vi kunnat effektivisera den manuella klassificeringsprocessen genom att låta endast en medarbetare klassificera en mening i materialet och ändå säkerställa god datakvalitet. I stället har nu två medarbetare behövt klassificera samma mening, för att därigenom upptäcka och korrigera skiljaktigheter mellan medarbetarna.

Med anledning av att anslagsvillkoren utgör en så brokig samling meningar undrar vi om den styrningen inte borde vara utformad på ett mer strukturerat sätt. I vissa fall har anslagsvillkoren varit så luddigt eller krångligt formulerade att de helt säkert har skapat onödigt administrativt arbete hos myndigheterna.

För att besvara frågan om hur väl en språkmodell fungerar för uppgiften att klassificera anslagsvillkor som styrande eller icke styrande kan vi säga att den finjusterade KB-BERT-modellen fungerar ganska bra, men att det finns utrymme för förbättring. Modellen klassificerar 85 procent av villkoren rätt, vilket är betydligt bättre än om den bara gissat på den vanligaste klassen, då den fått rätt i 46 procent av fallen. Modellen är bättre på att klassificera de klasser som är vanligt förekommande i träningsdatamängden än de klasser som är ovanligare.

Att modellen blir bättre på att hantera de klasser där flest klassificerade meningar finns representerade i träningsdatamängden är inget konstigt. I vidare studier skulle större fokus därför kunna ligga på att förbättra modellens förmåga att klassificera de mer sällan förekommande klasserna.

I ESV:s studie har vi upptäckt både för- och nackdelar med att undersöka i vilken utsträckning regeringen styr myndigheter med anslagsvillkor med hjälp av automatiskt klassificerad text. En fördel med metoden är att den möjliggör en kvantitativ analys av de data vi är intresserad av. Vi kan analysera större mängder data än vad vi hade kunnat göra med endast manuellt klassificerad data. En annan positiv aspekt är att metoden och analysen är reproducerbar. Om vi även för nästa år vill analysera anslagsvillkoren i regleringsbreven kan vi mata in (det förbearbetade)

textmaterialet i vår tränade språkmodell och strax därefter ha alla anslagsvillkor klassificerade. Språkmodellerna ger oss alltså möjlighet att enkelt och smidigt kunna klassificera nytt textmaterial utifrån samma frågeställning.

Nackdelen med att använda sig av automatisk klassificering jämfört med manuell klassificering är att den språkmodellbaserade klassificeraren inte presterar perfekt. Dock kan vi uppskatta en felmarginal med hjälp av en manuellt klassificerad testdatamängd. På så sätt kan vi också detektera modellens svagheter vad gäller specifika klasser. Det är dock inte säkert att en människa hade presterat så mycket bättre, och med största sannolikhet inte perfekt. I framtida studier vore det intressant att jämföra hur väl automatisk klassificering fungerar jämfört med mänsklig – att utvärdera också en människa mot testdatamängden. Detta som ett sätt att få en mer tolkningsbar felmarginal.

Referenser

Ahlbäck Öberg, Shirin och Wockelberg, Helena. (2020). ”Agency control or autonomy? Government steering of Swedish government agencies 2003–2017”, *International Public Management Journal*, DOI: 10.1080/10967494.2020.1799889.

Devlin, Jacob m.fl. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. *arXiv preprint. arXiv*:1810.04805.

ESV 2021:18. *Regeringens resultatstyrning av myndigheterna. En kartläggning av instruktioner och regleringsbrev under två decennier*.

Krippendorff, Klaus (1980, 2019), *Content Analysis. An introduction to its Methodology*.

Malmsten, Martin, Börjeson, Love och Haffenden, Chris (2020), “Playing with Words at the National Library of Sweden--Making a Swedish BERT.” *arXiv preprint arXiv*:2007.01658.

Neuendorf, Kimberly A. (2002), *The Content Analysis Guidebook*.

Sim, Julius och Wright, Chris C. (2005). “The kappa statistic in reliability studies: use, interpretation, and sample size requirements.” *Physical therapy*, vol. 85, nr 3/2005, s. 257–268.

Stollenwerk, Felix m.fl. (2022), “Annotated Job Ads with Named Entity Recognition.” [pdf], https://2022.sltc.se/papers/SLTC22_paper_3062.pdf. Hämtad 30 augusti 2023.

Wallerö, Emma. (2022). *Automatic Classification of Conditions for Grants in Appropriation Directions of Government Agencies*. [Masteruppsats, Uppsala Universitet]. DiVa. http://uu.diva-portal.org/smash/record.jsf?pid=diva2:1679811

Wei, Jason och Zou, Kai (2019), ”Eda: Easy data augmentation techniques for boosting performance on text classification tasks.” *arXiv preprint arXiv*:1901.11196.

Ordlista

Språkmodell

En språkmodell är kort och gott en algoritm eller ett program som kan hantera mänskligt språk genom att t.ex. klassificera, anonymisera, ordklasstagga eller transkribera sådan data åt dig. Det finns också språkmodeller som är generativa och alltså kan generera språkdata. Några exempel på arkitekturer för språkmodeller är t.ex. den regelbaserade Finite State Transducer-teknologin, Recurrent Neural Networks (RNNs), Bidirectional Encoder Representations from Transformers, och Generative Pre-trained Transformer (GPT).

Finjustering

Finjustering ("Fine-tuning" på engelska) är en teknik inom maskininlärning där en förtränad modell anpassas för en specifik uppgift. Detta görs genom att anpassa modellens vikter och parametrar med hjälp av data som är relevant för den specifika uppgiften, samtidigt som man bevarar den kunskap som modellen redan har fått genom sin ursprungliga träning.[26]

BERT

Bidirectional Encoder Representations from Transformers (BERT) är en typ av språkmodell som presenterades 2018 av forskare anställda på Google. BERT är baserad på en transformer-arkitektur. BERT är en transformermodell som är förtränad på stora mängder data. Modellen kan sedan finjusteras med ett extra output-lager.[27]

KB-BERT

KB-BERT är en svensk typ av BERT-modell som tränats av Kungliga biblioteket på 3497 miljoner ord.[28]

[26] Devlin, Jacob m.fl. (2018), *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint. arXiv:1810.04805.
[27] Devlin, Jacob m.fl. (2018.
[28]Malmsten, Martin, Börjeson, Love och Haffenden, Chris (2020), *Playing with Words at the National Library of Sweden--Making a Swedish BERT*. arXiv preprint arXiv:2007.01658.

ESV gör Sverige rikare

– Vi har kontroll på statens finanser, utvecklar ekonomistyrningen och granskar Sveriges EU-medel.
– Vi arbetar i nära samverkan med Regeringskansliet och myndigheterna.

Ekonomistyrningsverket
Drottninggatan 89
Box 45316
104 30 Stockholm
Tfn 08-690 43 00
www.esv.se